Increasing the Mixture Components of Non-Uniform HMM Structures Based on a Variational Bayesian Approach

Takatoshi Jitsuhiro, Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, “Keihanna Science City”, Kyoto 619-0288, Japan
{takatoshi.jitsuhiro, satoshi.nakamura}@atr.jp

Abstract

We propose using the Variational Bayesian (VB) approach for automatically creating non-uniform, context-dependent HMM topologies. Although the Maximum Likelihood (ML) criterion is generally used to create HMM topologies, it has an over-fitting problem. Recently, to avoid this problem, the VB approach has been applied to create acoustic models for speech recognition. We introduce the VB approach to the Successive State Splitting (SSS) algorithm, which can create both contextual and temporal variations for HMMs. Experimental results show that the proposed method can automatically create a more efficient model than the original method. Furthermore, we evaluated a method to increase the number of mixture components by using the VB approach and considering temporal structures. The VB approach obtained the best performance with a smaller number of mixture components in comparison with that obtained by using ML based methods.

1. Introduction

For creating acoustic models, phonetic decision tree clustering[1] is widely used as a method of generating tied-state structures. The Maximum Likelihood (ML) criterion is used to choose the phonetic question with which each state is split. However, owing to the nature of ML estimation, the ML criterion often causes a model that over-fits the training data. The likelihood value for training data increases as the number of parameters increases. Therefore, it is impossible to find the best model by using only the ML criterion. Methods based on the ML criterion require heuristic stop criteria, such as the total number of states.

To solve this problem, information criteria such as the Minimum Description Length (MDL) criterion and the Bayesian Information Criterion (BIC) have been introduced as splitting and stop criteria for creating context-dependent Hidden Markov Models (HMM). There are also some methods using phonetic decision tree clustering[3], or the Successive State Splitting (SSS) algorithm[4]. These methods continue to split states in order to improve the information criteria. Although they work well in practical terms, conventional information criteria require some assumptions, e.g., asymptotic normality, and they cannot exactly evaluate complicated models like neural networks, or HMMs, which cannot satisfy such assumptions.

In the field of machine learning, the Variational Bayesian (VB) method was proposed to avoid over-fitting by ML estimation[5]. Recently, some methods including the VB approach have been proposed for speech recognition. Decision tree clustering with the VB method was proposed[6], and Variational Bayesian GMMs were applied to speech recognition[8].

Latent variables are one of the key points in the VB framework. In [6], their method for making HMM structures does not require latent variables because the alignments of states are given. In [8], their models do not consider any temporal structures.

We propose an automatic topology creation method using the SSS algorithm with the Variational Bayesian method, which we call the VB-SSS algorithm, to estimate topologies more exactly. The SSS algorithm can create contextual and temporal variations. On the other hand, decision tree clustering can only create contextual variations. Furthermore, latent variables should be used in the SSS algorithm because the alignments of phonemes are fixed but those of states are not. Therefore, the occupancy probabilities of training samples should be considered by using latent variables to introduce the VB method into the SSS algorithm.

We also evaluate a method for increasing the number of mixture components by using the VB approach for topologies obtained by the VB-SSS algorithm. In [7], they evaluated two methods for making Gaussian mixture models. One sets the same number of Gaussians per state for all states, and selects an appropriate model by a VB objective function. The other determines the number of Gaussians for each state by splitting and merging Gaussians in each state by using the objective function. In [8], they made GMMs by decreasing the number of mixture components in each phoneme. Since the VB-SSS algorithm generates HMM structures with temporal structures, our proposed methods consider temporal structures to make mixture models by splitting Gaussians with the VB approach.

In Section 2, we present the VB-SSS algorithm. Next, a method for increasing the mixture components is described in Section 3. In Section 4, we evaluate the performance of our proposed methods by continuous speech recognition. Finally, we conclude our results in Section 5.

2. Variational Bayesian Approach for SSS Algorithm

2.1. Overview of VB-SSS

Figure 1 shows the flow of the VB-SSS algorithm. In this section, the VB-SSS algorithm[9] is explained in brief. First, the topology of an initial model is set and its parameters are estimated. Second, the prior parameters for each state are set. Next, the posterior parameters for each state are estimated, and the VB objective function, $F_m$, (see [5] for details) is calculated as the baseline energy.

After that, each kind of splitting is done in the same manner as with the ML-SSS algorithm[2]. For each splitting, after two new states are created, the posterior parameters are estimated, and the energy gains of both the contextual splitting and the temporal splitting, are calculated. Next, the state splitting with the maximum energy gain is selected. If there is no
2.2. Contextual and temporal splitting

The probability density of the HMM $\Theta$, which has $N_s$ states with one Gaussian distribution and $N_a$ transitions for each state for both contextual and temporal splitting, is

$$p(O|\Theta) = \prod_{t=1}^{T} \mathcal{N}(\alpha_t; \mu_{s_t}; S_{s_t}^{-1})a_{s_{t}r_{t+1}},$$

where $O = \{\alpha_1, \ldots, \alpha_T\}$ is a set of training samples, $s_t$ is the state number at time $t$, and $r_t$ is the transition arc number at time $t$. $\mu_{s_t}$ is a mean vector at $s_t$, $S_{s_t}^{-1}$ is a covariance matrix at $s_t$, and $a_{s_{t}r_{t+1}}$ is a transition probability. $S_{s_t}$ is referred to as a precision matrix and is defined as an inverse matrix of a covariance matrix. We use a diagonal matrix as a covariance matrix. The maximum of $N_a$ is $N_s$, and $N_a$ in this paper can be replaced by $N_s$. However, this splitting algorithm can use $N_a = 2$ only.

The probability for the complete data set to which the latent variables are introduced is

$$p(O, Z|\Theta) = \prod_{t=1}^{T} \prod_{i=1}^{N_s} \prod_{j=1}^{N_a} \mathcal{N}(\alpha_t; \mu_{i,t}; S_{i,t}^{-1})a_{ij}^{(t)}z_{ij}^{(t)},$$

where $Z = \{z_{ij}^{(t)}\}_{i=1,j=1}^{N_s,N_a,T}$ is the set of latent variables.

When the $i$th state with the HMM parameter $\Theta_i$ is split into the $i$th state and the $j$th state, the parameter $\Theta_j$ is estimated for the current splitting, the splitting criterion can be represented by using the objective function $F$ as follows.

$$\Delta F_m^{(n+1)} = F_m^{(n+1)}(\overline{\Theta}_i) - F_m^{(n)}(\Theta_i),$$

where $n$ is the iteration number.

2.3. Priors

We assume that the probability of parameters can be factorized as follows.

$$p(\Theta) = p(N_s, N_a) p(\alpha|N_s, N_a) p(S|N_s) p(\mu|S, N_s).$$

We also assume the prior of $\alpha = \{\alpha_{ij}\}_{i=1,j=1}^{N_s,N_a}$, $\alpha_{ij} \geq 0$, $\sum_{j'=1,j' \neq j}^{N_s} \alpha_{ij} = 1$, is a Dirichlet distribution, and the prior of $\{\mu, S\} = \{\{\mu_i\}_{i=1}^{N_s}, \{S_i\}_{i=1}^{N_s}\}$ is a normal-Gamma distribution.

$$p(\alpha|N_s, N_a) = \prod_{i=1}^{N_s} \prod_{j=1}^{N_a} \mathcal{N}(\alpha_{ij}; \nu_{ij}; \xi_{ij}^{-1} b_{ij}^{-1}) \mathcal{G}(s_{ij}; \eta_0/2, b_{ijk}/2),$$

where $D$ is the order of parameters. $\mu_{ik}$ and $s_{ik}$ are the $k$th elements of $\mu_{s_t}$ and $S_{s_t}$, respectively. $\mathcal{N}(\cdot)$ is the Gaussian distribution, and $\mathcal{G}(\cdot)$ is the Gamma distribution. $\phi_0$, $\nu_{0k}$, $\xi_0$, $\eta_0$, and $b_{0k}$ are prior parameters. The definition of the Gamma distribution $\mathcal{G}(s; \eta, \lambda) = \frac{s^{\eta-1}}{\Gamma(\eta)} \exp(-\lambda s)$, where $\Gamma(\cdot)$ is the Gamma function.

2.4. Posteriors

The posterior probability densities can be derived from the Variational Bayesian EM algorithm.

$$q(\alpha|O, N_s, N_a) = \prod_{i=1}^{N_s} \prod_{j=1}^{N_a} \mathcal{N}(\alpha_{ij}; \nu_{ij}; \xi_{ij}^{-1} b_{ij}^{-1}) \mathcal{G}(s_{ij}; \eta_0/2, b_{ijk}/2),$$

$$\phi_{ij} = \phi_0 + \bar{N}_{ij}, \quad \bar{N}_{ij} = \sum_{t=1}^{T} \bar{z}_{ij}^t, \quad \bar{z}_{ij}^t = <z_{ij}^t >_{q(Z)},$$

$$q(\mu, S|O, N_a) = \prod_{i=1}^{N_s} \prod_{k=1}^{D} \mathcal{N}(\mu_{ik}; \nu_{ik}; \xi_{ik}^{-1} b_{ik}^{-1}) \mathcal{G}(s_{ik}; \eta_0/2, b_{ik}/2),$$

$$\nu_{ik} = \bar{N}_{ik} \xi_{ik}/\bar{N}_{i} \xi_0, \quad \xi_{ik} = \xi_0 + \bar{N}_i, \quad \eta_i = \eta_0 + \bar{N}_i,$n$$

$$b_{ik} = b_{0k} + \bar{e}_{ik}, \quad \bar{e}_{ik} = (\bar{s}_{ik} - \nu_{0k})^2,$$

where $\bar{\alpha}_i = \frac{1}{N_s} \sum_{t=1}^{T} \overline{z}_{i}^{t} \alpha_{ij}$, $\bar{\xi}_{ik} = \frac{1}{N_s} \sum_{t=1}^{T} \overline{z}_{i}^{t} (\nu_{1k} - \bar{s}_{ik})^2$.

The variational posterior probability of latent variables is also derived in the same manner as the unknown parameters. $F_m$ can be derived from these priors and posteriors.

3. Increasing Mixture Components by Using VB Approach

3.1. Splitting mixture method

After topologies are obtained by the VB-SSS algorithm, the number of mixture components is increased by the following algorithm based on the VB approach. In this algorithm, the number of mixture components is estimated for each phone.
[Splitting mixture method]

1. Set an initial model obtained by topology training. $M^{(0)} = 1, n = 0$.
2. Calculate the objective function $F_m^{(n)}$.
3. Iterate the following steps for each phoneme.
   (a) Split each distribution into two distributions in each state. $M^{(n+1)} = 2M^{(n)}$, (Fig. 2)
   (b) Estimate posterior distributions, and calculate the objective function $F_m^{(n+1)}$, repeatedly.
   (c) Stop splitting when $\Delta F_m^{(n+1)} = F_m^{(n+1)} - F_m^{(n)}$ is a negative number. Otherwise, $n = n + 1$, and go to 3a.

3.2. VB approach for increasing mixture components

In [7] and [8], they estimated the number of mixture components for each state because their methods are the same as those used for GMMs. On the other hand, the VB-SSS algorithm estimates model structures by considering the transition probabilities using the forward-backward algorithm. Therefore, our proposed method estimates the number of mixture components by using the forward-backward algorithm.

Gaussian mixture HMMs can be represented as follows.

$$p(O|\Theta) = \prod_{t=1}^{T} \left\{ \sum_{k=1}^{M_t} \omega_{t,k} N(\mathbf{x}_t; \mu_{t,k}, \Sigma_{t,k}^{-1}) \right\} \alpha_{s_{t-1}t+1}$$ \hspace{1cm} (5)

where $s_t$ is the state index at time $t$. $\{w_{t,k}\}^{M_t}_{k=1}$ is a set of mixture weights for state $t$. $\mathbf{\mu}_{t,k}$ is a mean vector, and $\mathbf{\Sigma}_{t,k}$ is a precision matrix. $r_t$ is an arc index at time $t$. $\{\alpha_{ij}\}^{N_{t-1}}_{j=1}$ is a set of transition probabilities.

The priors and posteriors for transition probabilities, mean vectors, and precision matrices can be defined as the same as those of the VB-SSS algorithm. $p(\alpha|N_s, N_\alpha) = \prod_{t}^{T} \mathcal{D}(\alpha_{s_{t-1}t+1}; \phi_\alpha)$ is for transition probabilities, and $p(\mu, \Sigma|N_s, M_t) = \prod_{t}^{T} \mathcal{N}(\mu_{t,k}; \nu_0, \Sigma_{t,k}^{-1}) \mathcal{G}(\Sigma_{t,k}; \nu_0/2, b_{\Sigma}/2)$ is for mean vectors and precision matrices. For mixture weights, a Dirichlet distribution can be used.

$$p(w|N_s, M_t) = \prod_{t}^{T} \mathcal{D}(w_{t,k}; \rho_0)$$

where $\rho_0$ is a prior parameter. The posterior probabilities for these probabilities and the VB objective function including mixture components can be derived.

For recognition, predictive posterior probabilities are used for the Bayesian approach.

$$p(x|m, O) = \prod_{t=1}^{T} \int p(x_t|\theta_{m_{s, t+1}}, m, O) p(\theta_{m_{s, t+1}}, m, O) d\theta_{m_{s, t+1}}$$

where $x = \{x_1, \ldots, x_T\}$ is a set of test data. The posterior probability is approximated by the variational posterior probability.

$$p(\Theta|m, O) \propto q(\{a_{ij}\}^{N_{t-1}}_{j=1}|m) \prod_{t}^{T} q(\{w_{t,k}\}^{M_t}_{k=1}|m) q(\mathbf{\mu}_{t,k}, \mathbf{\Sigma}_{t,k}|m)$$

$$p(x|m, O) \cong \sum_{t=1}^{T} < a_{ij} > q(\{a_{ij}\}^{N_{t-1}}_{j=1}|m)$$

$$< a_{ij} > q(\{a_{ij}\}^{N_{t-1}}_{j=1}|m) = \phi_{ij} / \sum_{j'=1}^{N} \phi_{ij'}$$

$$< w_{t,k} > q(\{w_{t,k}\}^{M_t}_{k=1}|m) = \rho_{tk} / \sum_{k'} \rho_{tk'}$$

$$< N(\mathbf{x}_t; \mu_{t,k}, \mathbf{\Sigma}_{t,k}^{-1}) > q(\mu_{t,k}, \Sigma_{t,k}|m) = \sum_{i=1}^{D} T(x_{ti}; \nu_{tk}, \Sigma_{tk}), f_{tk} = \eta_{tk}, \sigma_{tk} = b_{\Sigma} (\xi_{tk} + 1)/\xi_{tk} f_{tk}$$

$T$ is a Student-t distribution.

4. Experiments

4.1. Experimental conditions

In this section, we evaluated our proposed method by means of conventional continuous speech recognition. We compared our proposed method, the VB-SSS, to the ML-SSS algorithms. For the acoustic training set, we used Japanese dialog speech from the ATR travel arrangement task (TRA) database uttered by 166 males. The total speech period was 2.1 hours. We have already obtained more than 80% word accuracy for this task by using 5-hour training data with males and females. In this paper, absolute word accuracy rates were less than 80%. Since our proposed method needs a lot of time to create models due to problems of both algorithms and implementation, we could only use the smaller amount of training data. We will solve this problem in the near future.

For testing, we used dialog speech including 213 sentences from the TRA database uttered by a different set of 17 males. In this paper, we used the VB approach only for the splitting and stopping criteria. Multi-class composite bigram models [10] were used, and the vocabulary size was 5,000. The sampling frequency was 16 kHz, the frame length was 20 ms, and the frame shift was 10 ms. We used 12-order MFCC, ∆MFCC, and ∆ log power as feature parameters. Cepstrum mean subtraction was applied to each utterance. We used 26 kinds of phonemes and one silence. Three states were used as the initial model for each phoneme. One Gaussian distribution for each state was used during topology training. A silence model with three states was built separately from the phoneme models. For the acoustic training set, we used Japanese dialog speech including 166 males. The total speech period was 2.1 hours. We have compared our proposed method by means of conventional continuous speech recognition. We compared our proposed method. For the acoustic training set, we used Japanese dialog speech from the ATR travel arrangement task (TRA) database uttered by 166 males. The total speech period was 2.1 hours. We have already obtained more than 80% word accuracy for this task by using 5-hour training data with males and females. In this paper, absolute word accuracy rates were less than 80%. Since our proposed method needs a lot of time to create models due to problems of both algorithms and implementation, we could only use the smaller amount of training data. We will solve this problem in the near future.
4.2. Experimental results

Figure 3 shows the results by using the single Gaussian models. The VB-SSS algorithm obtained almost the same performance as the ML-SSS algorithm, but a smaller size model was obtained.

Figure 4 shows the results by using the splitting mixture method. Furthermore, Table 1 shows the total number of mixture components by the best model of the baseline and the models by using the VB approach with several values of prior parameter, \( \rho_0 \). They show that the VB approach obtained the best performance with a smaller number of Gaussians in comparison with that obtained by using the ML based method.

Table 1: Total number of mixture components and word accuracy (WA) rates

<table>
<thead>
<tr>
<th># of mixture components</th>
<th>WA[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 mixtures/state</td>
<td>75.96</td>
</tr>
</tbody>
</table>

\( \rho_0 = 0.001, 0.01, 0.1, 1.0 \)

\( \# \text{ of states} = 1,400 \).

5. Conclusion

We proposed using the Variational Bayesian approach for automatically creating non-uniform, context-dependent HMM topologies. We introduced the VB approach to the SSS algorithm to create contextual and temporal variations for HMMs and then defined posterior probability densities and free energy functionals. We evaluated the proposed method for word-based continuous speech recognition. With about 60% of the ML-SSS states, the VB-SSS achieved comparable performance. Furthermore, we evaluated a method for increasing the number of mixture components by using the VB approach. Experimental results showed that the VB approach obtained the best performance with a smaller number of Gaussians in comparison with that obtained by using the ML based method.

6. Acknowledgment

This research was supported in part by the National Institute of Information and Communications Technology.

7. References